COMP41680 Assignment 2

Deadline: Monday 27th April 2020

Overview:

The objective of this assignment is to scrape consumer reviews from a set of web pages and to evaluate the performance of text classification algorithms on the data. The reviews have been divided into seven categories here:

http://mlg.ucd.ie/modules/yalp

Each review has a star rating. For this assignment, we will assume that 1-star to 3-star reviews are "negative", and 4-star to 5-star reviews are "positive".

The assignment should be implemented as a single Jupyter Notebook (not a script). Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results.

Tasks:

In this assignment you should complete all of the following tasks:

1. Select **three** review categories of your choice. Scrape all reviews for each category and store them as three separate datasets. For each review, you should store the review text and a class label (i.e. whether the review is "positive" or "negative").
2. For each of the three category datasets:
    a. From the reviews in this category, apply appropriate preprocessing steps to create a numeric representation of the data, suitable for classification.
    b. Build a classification model to distinguish between "positive" and "negative" reviews using **one** of the following classifiers:

        Naive Bayes, Logistic Regression, Random Forests

    c. Test the predictions of the classification model using an appropriate evaluation strategy. Report and discuss the evaluation results in your notebook.
3. Evaluate the performance of each of your three classification models when applied to data from the other two selected categories. That is, for the selected categories (A,B,C), run the experiments:
    a. Train a classification model on the data from "Category A". Evaluate its performance on data from "Category B" and data from "Category C".
    b. Train a classification model on the data from "Category B". Evaluate its performance on data from "Category A" and data from "Category C".
    c. Train a classification model on the data from "Category C". Evaluate its performance on data from "Category A" and data from "Category B".

Guidelines:

- The assignment should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- For the assignment, only these third-party packages can be used: NumPy, Pandas, Scikit-learn, NLTK, Gensim, SciPy, Requests, BeautifulSoup, Matplotlib, Seaborn.
- Submit your assignment via the COMP41680 Brightspace page. Your submission should be in the form of a single ZIP file containing the notebook (i.e. IPYNB file) and your data.
- In the notebook please state your student number only.
- Hard deadline: Submit by the end of Monday 27th April 2020
    - 1-5 days late: 10% deduction from overall mark
    - 6-10 days late: 20% deduction from overall mark
    - No assignments accepted after 10 days without extenuating circumstances approval and/or medical certificate.